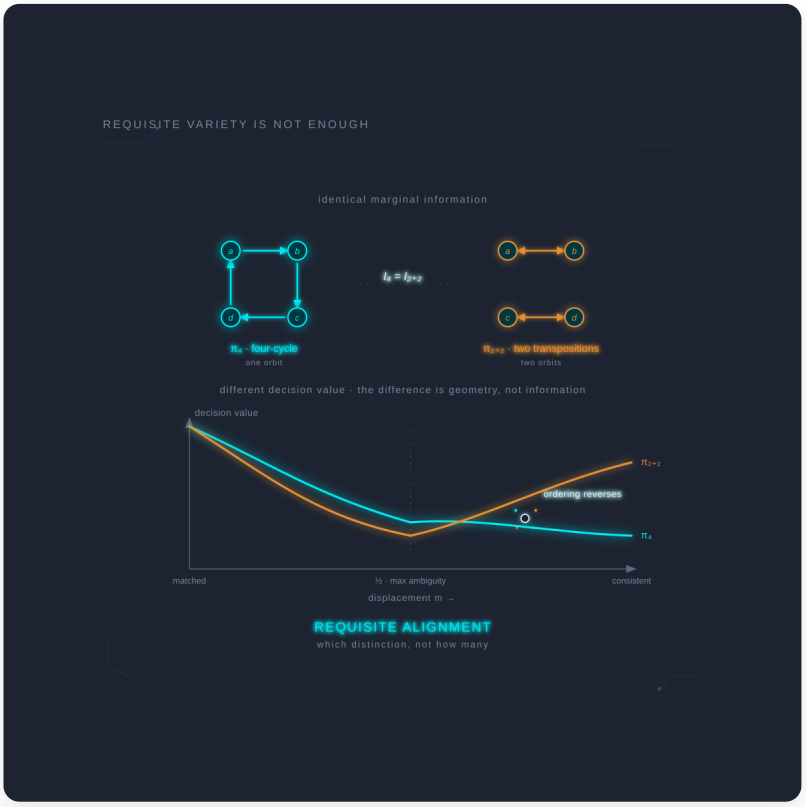

Requisite Alignment

The Geometry of Informative Feedback in a Solved Control Problem

Paper XXVII in the Governance as Engineering series

Asks whether dimensional sufficiency is enough. In a single exactly-solved partially-observed control problem, a fully informative feedback signal that evaluates a systematically wrong target loses substantial decision value — and two displacements of identical information content but different geometric structure impose different, oppositely-ordered costs. The result isolates a second constraint orthogonal to requisite variety: the distinction an observable resolves must be aligned with the distinction the controller's action turns on, and alignment has geometry. Paper XXVII in the Governance as Engineering series.

Björn Kenneth Holmström

July 2026

https://bjornkennethholmstrom.org/working-papers/requisite-alignment

Abstract

The Governance as Engineering series diagnoses institutional failure as insufficient *requisite variety* — an observation architecture whose dimensionality falls below that of the disturbance environment it governs. This paper asks whether dimensional sufficiency is enough. Working in a single exactly-solved partially-observed control problem, we measure by preregistered dynamic programming the decision value of a scalar feedback signal — the amount a controller's expected loss falls because it can pay to observe rather than act blind — and then displace that signal so that it evaluates a systematically *wrong* target while remaining fully informative and correctly modelled by an optimal controller. A first study establishes a large, connected region of the parameter space in which purchased feedback has resolved positive value. A second study, restricted to that region, finds that displaced-target feedback loses substantial value: resolved, substantive attenuation in 486 of 592 tested conditions (82.1%), unanimous among all conditions whose numerics resolved, and provably not reducible to a reduction in channel reliability. The decisive result concerns the *geometry* of the displacement. Two displacement geometries differing only in how the wrong targets are arranged across the class space — a connected four-cycle and two disjoint transpositions — impose different value costs, and the ordering reverses as displacement deepens: at low displacement the four-cycle preserves more value, at higher displacement the transpositions do. The two geometries are identical in variety by construction — same output cardinality, same reliability, same displacement probability, same likelihood multiset — so the difference is not a variety effect. We name it **requisite alignment**: for an observable to have decision value it is not enough that it carry sufficient variety and transmit it without compression — the distinction it resolves must be aligned with the distinction the controller's action turns on, and alignment has geometry (which distinction, not how many). The worst case is not maximal displacement but maximal *ambiguity* (intermediate displacement), because a consistently wrong signal can be partially inverted while an unpredictable one cannot. The study establishes that this failure mode is structurally real and geometrically structured in at least one exactly-analysable case; it does not estimate its prevalence in any governance system, and a registered possibility — that displaced feedback might exceed matched feedback in value — was permitted by the design but did not occur. This is the series' first computational-mechanistic paper, and its result refines the Goodhart–Ashby synthesis: requisite variety is necessary but not sufficient, and requisite alignment is a second, independently binding, geometrically structured constraint.

Keywords: requisite alignment, requisite variety, Goodhart–Ashby synthesis, value of information, partially observed control, feedback misalignment, observation architecture, Governance as Engineering.

Key claims (tiered)

- **[R]** The displaced feedback channel is not representable by any scalar reliability q: for m ∈ (0,1){½} it induces three distinct likelihood levels, and at m = ½ a class partition {a, π(a)} vs the rest, neither reproducible by a two-level q-channel. (§3.4, V5)
- **[R]** At m = 0 the displaced solve reduces exactly to the matched Gate-1 solve; dominance $A_\pi \geq 0$ holds; exact-inert planes remain $A_\pi = 0$ for all m and both geometries. (§3.6, V1–V3)
- **[R]** Under a uniform belief, the two registered geometries π_4 and π_{2+2} carry identical single-observation marginal mutual information about the state at every displacement level, because their likelihood vectors are relabelings of one multiset. The equality is prior-dependent and does not extend to the non-uniform beliefs the controller occupies after t = 0. (§3.4)
- **[R]** The geometry contrast holds the channel's *variety* fixed by construction — identical output cardinality, q, m, and likelihood multiset — so no account in terms of how many distinctions the channel can make separates π_4 from π_{2+2}. (§3.4, §5.2)
- **[open, registered]** Whether the geometry effect is additionally free of any information-quantity explanation along realized belief trajectories is untested. See RP-XXVII-1 (§5.2).
- **[R]** For K = 4 the fixed-point-free permutations comprise exactly two conjugacy classes (four-cycle, double-transposition); under the model's class-relabel symmetry one representative of each is exhaustive. (§3.4)
- **[R, numerical with envelope]** Over the frozen 74-cell panel, 486/592 conditions (82.1%) show resolved substantive attenuation, 0 enrichment, 0 practical equivalence, 106 non-convergent; attenuation is unanimous (486/486) among resolved conditions. (§4.2)
- **[R, numerical with envelope]** Geometry dependence is established by the registered criterion (≥ 2 adjacent qualified cells at a common m): the geometry contrast clears the substantive threshold on five adjacent pairs, and its sign reverses with displacement probability. (§4.4)
- **[H → R-in-model]** The value profile is non-monotone in displacement, with the minimum at intermediate ambiguity (near m = ½) rather than at maximal displacement — a consequence of a consistent displacement being partially invertible where an ambiguous one is not. Registered as a possible shape in advance; observed. (§4.3, §5.4)
- **[IP]** Requisite alignment is proposed as a general constraint — a relationship between what an observable resolves and what an action depends on — orthogonal to requisite variety and to integration-without-compression. The *concept* is general; the *evidence* is one solved instance. (§2, §5.6)

- **[H]** For governance: expanding observational dimensionality is necessary but incomplete; a system can fail by faithfully observing high-dimensional but misaligned distinctions, and the geometry of that misalignment — not only its extent — conditions the cost. (§5.6)

Registered negative (retained, not suppressed): the design permitted enrichment ($A_\pi > A_{matched}$ via Blackwell-incomparability of the displaced and matched channels); no resolved condition exhibited it on this panel. This constrains the enrichment hypothesis for this model, losses, and parameterisation; it does not establish that enrichment cannot occur under others. (§4.5, §5.3)

---

1. Introduction

The Goodhart–Ashby synthesis (Paper VI) holds that objective functions are observation architectures: a governance system whose value function projects a high-variety environment onto too few dimensions will, as a matter of formal necessity, exclude the disturbance dimensions that eventually destabilise it. This is the structural core on which the series' diagnosis rests. It follows from Ashby's Law of Requisite Variety and Shannon's channel capacity theorem, and it has organised the series' diagnosis across two dozen papers and twenty-one cases: the failures recur because they are the observable signatures of architectures operating below requisite variety.

Requisite variety, as the series has used it, is a claim about *dimensionality* — the number of independent dimensions along which a governance system can perceive its environment. A system has variety when it can register the clinical complexity healthcare excludes, the distributional consequences central banks exclude, the systemic patterns courts exclude. The prescription that follows is to raise the effective dimensionality of the observation channel: perceive more of the disturbance environment, along more independent axes, at each relevant scale.

This paper asks a question the dimensional framing leaves open. Suppose an observation channel *does* carry sufficient variety — it registers a genuine, independent dimension of the environment, at full information — but that dimension is the *wrong one for the action at hand*. It reports a real distinction, reliably, yet not the distinction the controller's intervention actually turns on. Does requisite variety, in the dimensional sense, guarantee that such a channel has decision value? Or is there a second constraint, orthogonal to dimensionality, that a channel must also satisfy?

We answer this not with an architectural argument but by solving the control problem exactly. We take a single, fully specified partially-observed decision problem — a controller facing four hidden disturbance classes, choosing countermeasures under a fixed loss structure, with the option to purchase a scalar feedback signal — and compute, by preregistered dynamic programming, the exact value of that feedback: how much a controller's expected loss falls because it can pay to *attend*, to observe rather than act blind. This is the series' first computational-mechanistic paper. Its object is unchanged — the value of an observable to a controller facing disturbance variety — but where the earlier papers establish structural necessity by [R]-tier proof about architectures, this paper measures a specific value surface by exact solution, with every claim tiered, every prediction frozen before code, and every numerical result carried with an explicit convergence envelope or censored.

The first study (Gate 1) establishes *where* purchased feedback has decision value at all, across the full parameter space of cue reliability, disturbance persistence, channel reliability, and acquisition cost. It finds a large, connected region in which attention is worth its cost — establishing that the quantity we are measuring is real and substantial before we perturb it. The second study (Gate 2) then displaces the feedback: it keeps the signal fully informative, correctly modelled by an optimal controller, and provably not merely noisier — but points it at a systematically *wrong target*, so that it evaluates a different class than the one the controller acted on. The displacement is parameterised by a probability m and, critically, by the *geometry* of the target permutation — whether the wrong targets form one connected cycle through the class space or two disjoint swaps.

The result is that requisite variety is not sufficient. A signal that carries information an optimal controller cannot recover by treating it as mere noise — one provably not reducible to a change in channel reliability — loses substantial decision value when it is aligned to the wrong distinction: resolved, substantively meaningful attenuation in 82% of tested conditions, and unanimous attenuation among every condition whose numerics resolved. The cleanest form of the claim comes from the *geometry* of the misalignment. Two displacements of equal probability and identical channel variety but different geometric structure — whether the wrong targets form one connected cycle through the class space or two disjoint swaps — impose different value costs, and, across the range of displacement, oppositely ordered ones: at low displacement a connected 4-cycle preserves more value than two disjoint swaps, and by higher displacement the ordering reverses. The two channels have the same output cardinality, the same reliability, the same displacement probability, and the same likelihood multiset; they differ only in which latent class each likelihood level is attached to. No account of how many distinctions the channel can make can therefore separate them, and the difference in value is an effect of *which* distinction the signal resolves relative to the controller's action. Whether it is additionally separable from any information-quantity account is a narrower question, registered as an open test in §5.2. The value of an informative signal depends on the geometry of the latent distinction it reports, not only on how reliably, how often, or along how many dimensions it reports.

We name this second constraint **requisite alignment**: for an observable to have decision value, it is not enough that it carry sufficient variety (Ashby) and transmit that variety without compression (the series' integration-without-compression property); the distinction the observable resolves must be *aligned with the distinction the controller's action turns on*. Requisite variety is a claim about the dimensionality of perception. Requisite alignment is a claim about its geometry — which distinction, not how many. This paper shows, in a solved model, that alignment is a binding constraint independent of variety, and that it has structure: alignment can be lost by degrees and along different geometries, with the cost of misalignment depending on both.

2. Relation to the series: a refinement of the synthesis

2.1 What the synthesis says, and what it leaves open

The Goodhart–Ashby synthesis identifies a *dimensional* failure. An objective function — equivalently, an observation architecture — that projects a high-variety environment onto too few dimensions will omit disturbance dimensions that accumulate as externalities until they force crisis. The corrective, in the series' terms, is to expand the effective dimensionality of the observation channel: the healthcare system must perceive clinical complexity, not only administrative throughput; the central bank must perceive distributional consequences, not only aggregate stability. The whole diagnostic machinery of the series — the averaging problem (Paper I), the fractality requirement (Paper II), constitutional unobservability (Paper III), commons collapse under low-dimensional channels (Paper IV) — turns on the *number* of independent dimensions the architecture can register relative to the environment it faces.

The series does contain one concept adjacent to alignment: *integration without compression* (Paper VII) — coordination mechanisms that carry a signal from the scale of observation to the scale of action without destroying it. But that property concerns transmission *fidelity*: it asks whether a genuine signal survives its journey through the architecture, not whether the signal was pointed at the right distinction to begin with. A perfectly transmitted signal about the wrong distinction is a failure that integration-without-compression does not name.

This is the gap the present paper fills. Ashby's Law states that a controller must command variety at least equal to the variety of disturbances it must reject. It is a theorem about *sufficiency of dimension*. It is silent on *correspondence*: whether the variety the controller commands is the variety of the *relevant* distinction. Two observation channels can have identical dimensionality, identical mutual information with the environment, and identical transmission fidelity, and yet one can be worth a great deal to a given controller and the other worth nothing — because one resolves the distinction its action depends on and the other resolves an equally rich but misaligned one.

2.2 What the solved model adds

The value of solving the model exactly, rather than arguing architecturally, is that it lets us separate alignment from variety — an experiment the dimensional framing cannot express. The displaced feedback channel in Gate 2 is constructed so that its marginal reliability parameter is unchanged and it remains provably non-reducible to a change in channel reliability: it is not a lower-variety channel

in the dimensional sense, and it is not a noisier q. It carries information about a real class distinction. What changes is only *which* distinction — whether the signal evaluates the class the controller acted on or a permuted image of it. The cleanest separation comes from the two permutation geometries, which are identical in variety by construction — same output cardinality, q, m, and likelihood multiset — yet differ in decision value: because variety is held exactly fixed, the value difference isolates alignment as an axis distinct from variety. (Their single-observation marginal information is also equal, but only under a uniform prior; §3.4.). Whether it is additionally distinct from information *quantity* along a trajectory is an open question registered in §5.2, since the uniform-prior equality does not extend to the beliefs the controller occupies after the first observation. (The displacement probability m does also change the channel's total mutual information at interior values — the value cost as m rises is therefore partly confounded with an information-quantity effect, which is why the geometry contrast, not the raw m-attenuation, is the load-bearing evidence; §5.2 treats this carefully.)

Three findings sharpen the synthesis. First, requisite variety in the dimensional sense is necessary but not sufficient: a correctly-modelled signal, non-reducible to added noise, loses most of its value under misalignment. Second, alignment is continuous, not binary — value degrades by degrees as the displacement probability rises, and non-monotonically, recovering somewhat at full displacement where the signal becomes a clean report about a fixed (if wrong) target. Third, and most consequential for the series, alignment has *geometry*: at identical channel variety, the cost of misalignment depends on the structure of the mismatch — how the wrong targets are arranged across the state space — with a reversal in which geometry is less costly as the displacement deepens. The third finding is the one that carries the argument, because it is the one that holds the channel's variety exactly fixed; the first two are consistent with, but do not by themselves prove, a constraint distinct from variety (§5.2).

The third finding is the one that most extends the framework, because it shows that "which distinction" is not a single bit of correspondence but a structured object. The series' dimensional account has no vocabulary for it: two channels of equal dimensionality and equal misalignment magnitude can still differ in value because of how their misalignment is arranged geometrically. Requisite alignment is therefore not a scalar corrective to be bolted onto requisite variety; it is a second, geometrically structured constraint that a governance architecture must satisfy independently.

2.3 Requisite alignment, defined within the model

The refinement can be stated precisely in terms of the model's own objects. Fix the controller's action a and let z be the hidden class. Two partitions of the state are in play. The **action-relevant distinction** is the partition the loss depends on: under the registered loss the outcome turns on whether the countermeasure matches the class, i.e. on the indicator 1[z = a] — the loss rewards separating the acted class from the rest. The **signal-resolved distinction** is the partition the feedback is informative about: the likelihood P(match | z) separates the classes according to which one the comparator target evaluates. In the matched channel (m = 0) the signal is informative about 1[z = a] — it resolves exactly the distinction the loss turns on. Displacement moves the signal-resolved distinction onto 1[z = π(a)] (fully at m = 1) or a mixture (at intermediate m).

Within this model, requisite alignment is the correspondence between the signal-resolved distinction and the action-relevant distinction. A channel is aligned when the partition it is informative about agrees with the partition the loss depends on, and misaligned to the degree they diverge. Requisite variety asks whether the channel's partition is fine enough (whether it carries enough information about the state); requisite alignment asks whether that information is about the *right* partition — the one the action turns on. The two are independent: displacement moves the signal-resolved distinction away from the action-relevant one while leaving the channel's variety untouched, and decision value falls even though variety does not. The two registered geometries make the separation exact — they are built from a common likelihood multiset and differ only in which latent class each level is attached to (§3.4).

The correspondence admits a *partial* order but not a total one, and this is the crux. Two partitions can be compared by refinement — whether one is a coarsening of the other — but the signal-resolved and action-relevant partitions here are generally incomparable under refinement, and, crucially, alignment is **not** reducible to any information ordering between them. The two registered geometries are the proof: they are variety-equivalent, and information-equivalent under a uniform prior (§3.4), yet alignment-distinct (different decision value, with a sign reversal). No variety measure can separate them, by construction; alignment does. Requisite alignment is therefore a *finer* object than variety — it is the geometry of how the signal-resolved and action-relevant partitions differ, not the number of distinctions the signal can make. This model exhibits three properties of that correspondence, each established here [R, within-model]: (i) misalignment reduces decision value at fixed variety (the geometry contrast); (ii) alignment is graded — in the displacement probability m and in the geometry — and its cost is non-monotone, maximal at intermediate

ambiguity rather than at maximal displacement, because a consistently misaligned signal is partially invertible where an ambiguous one is not; (iii) equal degrees of misalignment with different geometric structure impose different, even oppositely-ordered, costs.

Scope [IP]. The definition and its three properties are stated for, and established in, this single finite control problem: K = 4, the registered loss, a binary action-coupled feedback channel, and a known permutation displacement. We conjecture that an analogous correspondence — between the distinction an observable resolves and the distinction an intervention turns on — is a binding constraint on the value of information in control problems of this kind, independent of and additional to requisite variety. That conjecture is in-principle: it is motivated by the model but not proved beyond it, and this paper establishes no necessary or sufficient conditions for it in the general case. What is established [R] is that the constraint is real, independent of channel variety, and geometrically structured in at least one exactly-solvable instance. Whether it is additionally independent of information quantity is registered as an open test (RP-XXVII-1, §5.2).

2.4 Why this is a governance result, at the scale of a single observer

The controller in this model is a single attending agent, and the motivating question — what is sustained attention worth? — arose from the series' Self sub-series, which applies the variety grammar at individual scale. It would be a mistake to read the paper as therefore *about* individual cognition rather than governance. The quantity measured, the value of a purchased observation to a controller facing disturbance variety, is a control quantity at any scale; the machinery is identical whether the controller is a mind, an institution, or a mechanism. What the single-observer setting buys is tractability: a governance architecture with many coupled observers cannot be solved exactly, but a single controller's feedback value can, and the constraint it reveals — requisite alignment — is scale-free in the same way requisite variety is. The result is a governance result demonstrated at the smallest scale at which the relevant control problem can be solved to the last decimal, and its implication scales upward: an institution can enlarge the dimensionality of its dashboards indefinitely and still govern badly if the distinctions those dashboards resolve are misaligned with the distinctions its interventions turn on — and the *geometry* of that misalignment, not only its extent, determines how badly.

2.5 Related theory

The claim that the value of information depends on more than its quantity is not itself new, and the paper's novelty must be located precisely against three established bodies of work. **Blackwell's comparison of experiments** (Blackwell, 1951, 1953) established that one signal is more valuable than another across *all* decision problems if and only if the second is a garbling of the first — a partial order strictly coarser than mutual-information dominance, and the exact tool with which we establish (§3.5) that the displaced and matched channels are incomparable, so that neither attenuation nor enrichment is guaranteed a priori. **Classical value-of-information theory** (Howard, 1966; and the decision-theoretic tradition following it) makes the value of a signal explicitly dependent on the prior, the payoff structure, and the signal's diagnostic relationship to the available decisions — so that "information quantity does not determine decision value" is, in that literature, a settled starting point rather than a discovery. And the **Conant–Ashby good regulator theorem** (Conant & Ashby, 1970) already makes *correspondence* between regulator and regulated system central to regulation, not raw variety — the regulator must be a homomorphic image of the system it controls.

Against this backdrop the contribution is not the general claim that information value is context-dependent. It is specific and constructive: (i) a preregistered, exactly-derived family of action-coupled *target-displacement* channels, analytically shown to be non-representable by any scalar reliability; (ii) a solved sequential POMDP value surface for those channels; (iii) the comparison of two *non-conjugate* permutation geometries at matched variety and matched uniform-prior information content; (iv) the resulting non-monotone value profile with partial recovery at full displacement; and (v) a resolved, spatially replicated geometry-by-displacement crossover. The requisite-alignment framing then integrates these into the Governance as Engineering programme, where the pertinent question is not whether information value is context-dependent in general, but which *structural* property of an observation channel — beyond variety and beyond transmission fidelity — a governance architecture must satisfy. The good regulator theorem locates that property in the regulator's model of the system; requisite alignment locates a companion property in the observation channel's target, and shows in a solved case that it is geometrically structured. Relatedly, the problem of choosing *what* to observe is the subject of POMDP sensor selection and observation design (Kaelbling, Littman & Cassandra, 1998, for the POMDP setting); this paper differs in fixing the channel's variety and its uniform-prior information content, and varying only its action-relative target geometry.

3. Methods

3.1 The control problem

We study a finite-horizon partially observed control problem in which a controller repeatedly faces one of **K = 4** hidden disturbance classes and must choose, at each step, either a null action or a class-specific countermeasure. The loss structure is fixed throughout: a correct countermeasure incurs L_correct = 0 plus a fixed action cost κ = 0.3; the null action incurs L_null = 1.0; a wrong countermeasure incurs L_wrong = 2.0. Because L_correct + κ = 0.3 < L_null < L_wrong, acting correctly is preferred to inaction, which is preferred to acting wrongly — the controller has a genuine identification problem, not merely a detection one.

The controller observes a contextual cue about the hidden class with reliability r (at r = 1 the cue reveals the class exactly; at lower r it is proportionally less informative). The disturbance process is Route B (hold-or-reset): with probability p the class and cue persist, and with probability 1 − p the system resets — a new class is drawn independently and a fresh cue emitted. Thus p is a retention coefficient and 1/(1 − p) the expected interval between resets; p = 0 is genuinely memoryless.

The object of study is a second, optional information channel: after taking a real countermeasure the controller may pay acquisition cost c to receive one **binary** feedback observation, approximately "was the chosen countermeasure correct?", with channel reliability q (at q = ½ the channel is uninformative). Feedback arrives after the current action and loss, so it can improve only future decisions, and purchasing it is optional. Writing J*_C1 for the minimum expected cumulative loss with the cue alone and J*_C1+C2 for the minimum when feedback may also be purchased, the quantity of interest is the **decision value of feedback**

> $$A = J^*_{C1} - J^*_{C1+C2} \geq 0,$$

nonnegative because the controller may always decline to purchase (the dominance property). A is the value of a controller's option to *attend* — to pay to observe — over and above the context it already has. The horizon is H = 64 throughout.

3.2 Solver and the two-study structure

Both studies use a single belief-grid backward-induction POMDP solver over the K = 4 belief simplex, with closed-form CDF-Kuhn (Freudenthal) interpolation on a regular simplex grid at refinement levels G = 40 → 54 → 64. (An alpha-vector representation was tried and rejected as intractable for K = 4; the Kuhn interpolation replaced a scipy-Delaunay implementation whose `find_simplex` cost was pathological at these grid sizes — ~28 s per call at G = 64 — after verifying the two interpolants converge to a common limit and that A, being a symmetric functional of the value functions, is exactly class-symmetric under Kuhn despite the interpolant itself not being permutation-equivariant.) Every claim carries an epistemic tier: **[R]** rigorous (structural, method-independent), **[IP]** in principle, **[H]** heuristic. Numerical results carry an explicit convergence envelope and are censored when unresolved.

The work proceeds as two preregistered gates. **Gate 1** establishes *where* purchased feedback has decision value at all — the sign question, $A > 0$ vs $A \approx 0$ — over the full four-dimensional parameter grid (r, p, q, c), each axis at seven levels (2,401 cells). **Gate 2** conditions on the resolved positive region and asks how that value changes when the feedback signal is **displaced** — when it evaluates a systematically different target than the intervention taken. Predictions, classification rules, and validation obligations were frozen before any solver was written, following the series' simulation-first discipline; both preregistrations and every numerical-method amendment are recorded with provenance hashes.

3.3 Gate 1 — the activation surface

Gate 1 brackets A with two independent bounds that share no Bellman core with the solver: a lower bound from a validated one-shot Monte-Carlo policy (a simultaneous 99% empirical-Bernstein certificate over 144 nondegenerate triples yielding 562 high-confidence-active cells), and an upper bound reducing to the 889 exact "known-answer" planes where $A = 0$ by structural certificate ($p = 0$ memoryless, $q = ½$ uninformative, or $r = 1$ certain). The solver then classifies every cell, refining the boundary band with the three-level Kuhn sequence and a dominance-respecting rule that separates resolved-active cells from exact-inert planes, from cells numerically indistinguishable from zero (zero-compatible), from genuinely non-convergent cells. A registered convergence amendment (a fixed absolute-stability tolerance $T_{abs} = 10^{-8}$, verified invariant across five orders of magnitude) corrected a rule defect that had mislabeled stably-zero cells as non-convergent.

The frozen Gate-1 surface is **1,121 active / 889 exact-inert / 277 zero-compatible / 114 non-convergent** cells, with zero dominance violations and the two registered anchors resolving as predicted (the high-value anchor active; the low-value anchor computing $A \approx 0$ to machine precision

at all three grid levels). A companion consistency guard confirmed all 288 non-plane inert cells are behaviourally inert (the optimal controller declines to purchase where feedback has no value).

3.4 Gate 2 — the displaced-feedback mechanism

Gate 2 modifies only the feedback observation likelihood. Given the controller's action a, a fixed, **known** permutation π of the class labels, and a displacement probability m, the comparator target on each purchased observation is drawn independently:

> $T = a$ with probability $(1 - m)$, $T = \pi(a)$ with probability m,

and the binary signal reports match with probability q if T equals the true class, else 1 − q. Integrating out T gives the displaced channel likelihood over the true class z [R]:

> $P(\text{match} \mid z = a) = q - m(2q - 1)$; $P(\text{match} \mid z = \pi(a)) = 1 - q + m(2q - 1)$; $P(\text{match} \mid z = \text{other}) = 1 - q$.

The controller knows π and m and updates optimally under this known ambiguity; it is not a deceived agent but one reading a **known but stochastically displaced, action-coupled sensor**. Crucially, this is not a rescaling of the reliability q. The matched q-channel has two likelihood levels (q at the acted class, 1 − q elsewhere); the displaced channel has **three** distinct levels for $m \in (0,1) \setminus \{½\}$, and at m = ½ it has two levels but partitions the classes as {a, π(a)} versus the rest — a partition no scalar q-channel can produce, since a q-channel always separates the acted class from all others [R]. The displacement therefore cannot be recovered by reading a reliability slice off the Gate-1 surface; the permutation acts on the comparison *target* in K-dimensional class space, upstream of the binary output, and that structure survives the collapse of the signal to a single bit.

Lemma (uniform-prior information invariance of the displacement) [R]. For fixed q and m, and for a controller belief that is uniform over the K classes, the single-observation mutual information between the true class and the binary signal is invariant across *all* fixed-point-free permutations π and all actions a. The likelihood vector P(match | z) is, for every such π and a, a relabeling of the same multiset $\{q - m(2q-1), 1 - q + m(2q-1), 1 - q, 1 - q\}$. Writing the mutual information at belief b as $I(b, L) = H(b \cdot L) - b \cdot H(L)$, both terms are symmetric functions of the likelihood multiset whenever b is exchangeable over the classes; the quantity therefore depends only on (q, m). In

particular the two registered geometries π_4 and π_{2+2} carry identical marginal information at every m under a uniform prior. (Verified across all nine K = 4 derangements and four actions; max deviation 3.3×10^{-16} bits.)

Scope, stated explicitly. The exchangeability condition is load-bearing and was left implicit in an earlier version of this lemma. Mutual information is a functional of the belief as well as the likelihood, and at non-uniform beliefs the two geometries are *not* informationally equivalent: they assign the same likelihood levels to different latent classes, so b·L and b·H(L) both differ. At q = 0.85, m = 0.75, per-observation MI differs between π_4 and π_{2+2} by up to 0.22 bits across the belief simplex — larger, in parts of the parameter range, than the channel's entire uniform-prior information content. A worked instance: at q = 0.7, m = 0.5, a = 2 and belief (0.05, 0.05, 0.45, 0.45), π_4 yields likelihoods (0.3, 0.5, 0.5, 0.3) and π_{2+2} yields (0.5, 0.5, 0.3, 0.3) — the same multiset — for MI of 0.0303 and 0.0112 bits respectively, against a common uniform-prior value of 0.0303 bits.

Since the controller's belief is uniform only at t = 0, the geometry contrast is conducted almost entirely at beliefs where the equality above does not hold. What the contrast holds fixed without qualification is the channel's *variety*: output cardinality, q, m, and the likelihood multiset are identical by construction. What it does not establish by construction is freedom from any information-quantity account along a trajectory. That question is treated in §5.2 and registered there as an open prediction rather than settled here.

Two permutation geometries are registered as co-primary. For K = 4 the fixed-point-free permutations fall into exactly two conjugacy classes: the connected **4-cycle π_4 = (1 2 3 4)** and the **double-transposition π_{2+2} = (1 2)(3 4)** [R, by enumeration]. Because the base model is symmetric under class relabeling, one representative of each class is exhaustive; π_4 and π_{2+2} are not conjugate to each other, so they are genuinely distinct channels rather than relabelings.

3.5 Estimand, panel, and classification (frozen)

The estimand is the **signed** advantage change $D_\pi(m) = A_\pi(m) - A_{matched}$, computed directly as $J^*_{matched} - J^*_\pi$ so that the common cue-only baseline J^*_{C1} cancels exactly and its interpolation error does not enter the envelope [R]. The direction is not presumed: outcomes are classified as attenuation, enrichment ($A_\pi > A_{matched}$), practical equivalence, or unresolved. No upper envelope $A_\pi \leq A_{matched}$ is assumed — the displaced and matched channels are Blackwell-incomparable (a garbling of the matched channel cannot distinguish $\pi(a)$ from the other wrong classes, so displacement can in principle redistribute rather than only destroy information), so enrichment is a registered possible finding rather than a defect [R].

The study runs on a **frozen 74-cell panel**, selected deterministically from the Gate-1 surface: the qualified interior (active cells with $A - \varepsilon_A > 10 \cdot \varepsilon_A$, i.e. magnitude trustworthy, not merely sign), stratified by tertiles of A with hash-ordered selection, the registered anchors and slices where they qualify, and one adjacency pair per tertile. The panel is a purposive sample for existence and heterogeneity, not prevalence. Each cell is solved for the matched baseline and for both geometries at $m \in \{0, 0.25, 0.5, 0.75, 1.0\}$ at all three grid levels.

Classification is interval-based and gated on convergence. A condition's D_π is classifiable only if **both** $J^*_{matched}$ and J^*_π independently converge (each fine refinement gap no larger than its coarse gap, or both below $T_{abs} = 10^{-8}$); otherwise the condition is censored as non-convergent regardless of how stable D_π appears. A converged effect is substantive only if it clears the registered minimum $\delta_{min} = 0.05 \cdot \text{median}(A \text{ over the panel}) = 0.648$ *after* its uncertainty: attenuation requires $D_\pi + \varepsilon_D \le -\delta_{min}$, enrichment $D_\pi - \varepsilon_D \ge +\delta_{min}$, practical equivalence $|D_\pi| + \varepsilon_D < \delta_{min}$. The $m = 0$ condition is a validation endpoint (D_π must be zero exactly) and is excluded from outcome fractions. Geometry dependence is registered as established only if the contrast $G(m) = D_{\pi_4} - D_{\pi_{2+2}}$ clears δ_{min} on at least two adjacent qualified cells at a common m, or on a slice.

3.6 Validation obligations

Eight structural obligations were required to pass before any scientific condition was interpreted: (V1) $m = 0$ reduces exactly to the matched Gate-1 solve; (V2) dominance $A_\pi \ge 0$, with a resolved parity violation requiring the interval $A_\pi + \varepsilon$ to clear zero; (V3) exact-inert planes remain $A_\pi = 0$ for all m and both geometries, computed through the channel code rather than by shortcut; (V4) the implemented likelihoods reproduce the three-level analytic form; (V5) non-representability by any scalar q; (V6) the conjugacy-symmetry spread across all permutations of a cycle type falls within the displaced-policy envelope; (V7) an independent exact $H = 2$ enumeration, sharing no interpolation core, reproduces the solver within envelope and confirms the *model* is exactly conjugacy-symmetric (so any solver-level spread is interpolation bias, not a model asymmetry); and (V8) an anchor-and-plane pilot before the interior run.

4. Results

4.1 Validation

All eight obligations passed (Table 1). The m = 0 identity held to machine precision (0 failures across the panel), confirming that the displaced solver reduces exactly to Gate 1 and that the modification is confined to displaced behaviour. Dominance held with zero parity violations. The exact-inert planes returned $A_\pi = 0$ to machine precision for all m and both geometries. The independent H = 2 enumeration gave a model conjugacy spread of 4×10^{-16} — the model is exactly permutation-symmetric — while the interpolated solver's conjugacy spread, though not machine-zero (the Kuhn interpolation is coordinate-order-dependent), shrank with refinement ($3.97 \times 10^{-2} \rightarrow 1.73 \times 10^{-2} \rightarrow 1.69 \times 10^{-3}$ at G = 18/27/40 on the anchor at m = ½) and sat roughly three orders of magnitude below the symmetry envelope. The single-representative-per-geometry reduction is therefore valid, and the solver spread is confirmed as interpolation bias rather than a broken model symmetry.

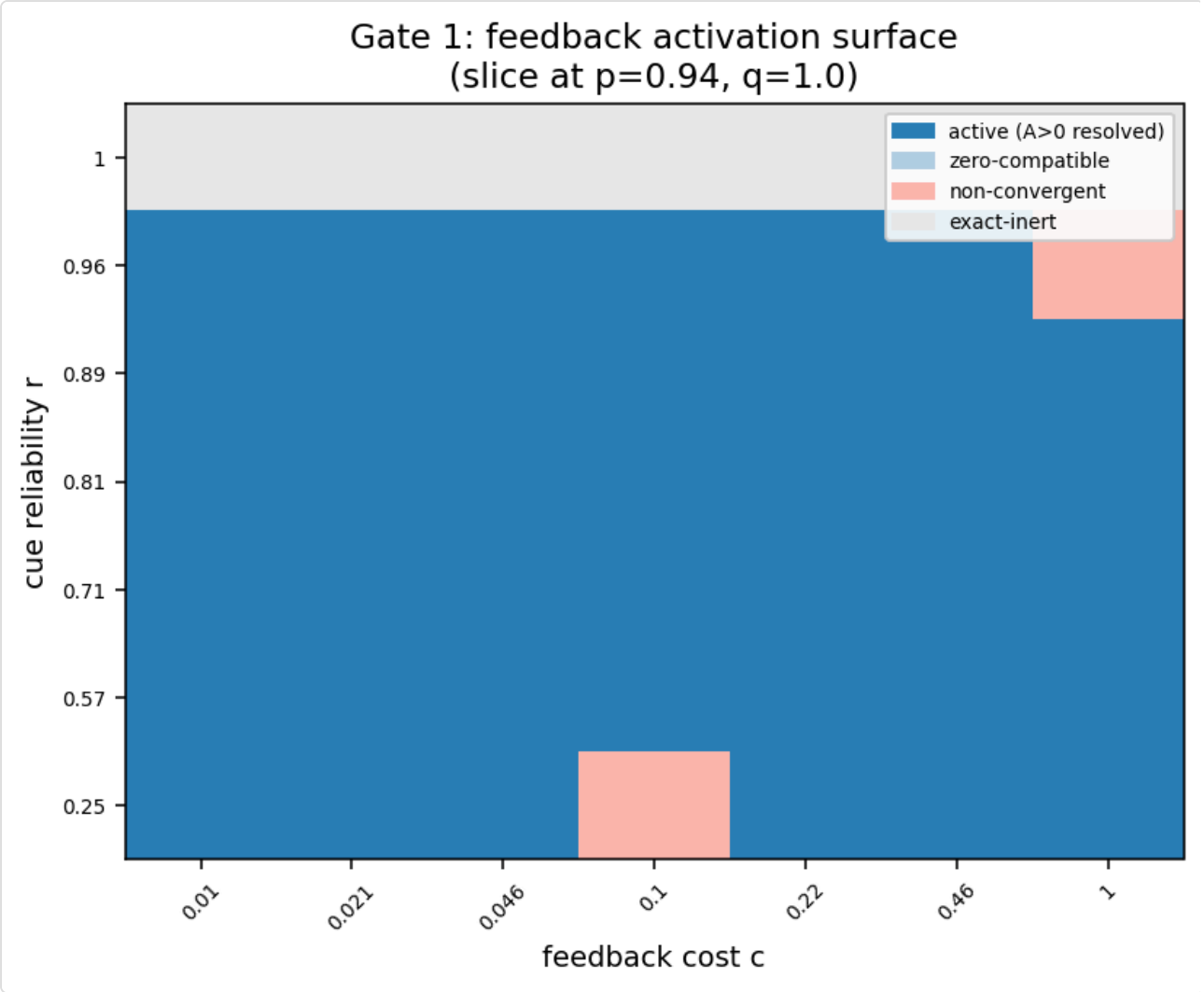

Figure 1. Gate-1 activation surface on the (cue reliability r × feedback cost c) slice at p = 0.94, q = 1.0. Purchased feedback has resolved positive value (active) across a connected high-reliability region; as acquisition cost rises the value falls below resolution (zero-compatible) or the cell is a structural plane (exact-inert). This establishes the baseline the displacement study conditions on: a substantial region where matched feedback is worth its cost.

Table 1. Validation obligations (V1–V8) and outcomes.

#	Obligation	Result
V1	m = 0 reduces exactly to the matched Gate-1 solve	PASS — 0 failures across the panel; displaced solver = matched baseline to machine precision at m = 0
V2	Dominance $A_\pi \geq 0$ (resolved parity requires interval $A_\pi + \varepsilon < 0$)	PASS — 0 parity violations; min margin $A_\pi + \varepsilon_D = +0.0014$
V3	Exact-inert planes remain $A_\pi = 0$ ∀m, both geometries	PASS — max $\lvert A_\pi \rvert = 0$ over the plane subset × geometries × m, computed through channel code
V4	Implemented likelihoods reproduce the analytic form	PASS — three levels for $m \in (0,1) \setminus \{½\}$, two-level partition at m = ½, matched two-level at m = 0; machine precision
V5	Non-representability by any scalar reliability q	PASS — instantiated at m = 0.25 (three levels) and m = ½ ({a, π(a)} vs rest)
V6	Conjugacy-symmetry spread within the displaced-policy envelope	PASS — solver spread shrinks with refinement (3.97e-2 → 1.73e-2 → 1.69e-3 at G = 18/27/40) and sits ~10^3× below σ_{sym}
V7	Independent exact H = 2 enumeration (no shared interpolation core)	PASS — model conjugacy spread 4.4e-16 (exactly symmetric); q = ½ cells A = 0
V8	Anchor-and-plane pilot before the interior run	PASS — S1 and V3 exact on pilot

4.2 The primary outcome: attenuation, unanimous where resolved

Over the eligible set of 592 conditions (74 cells × 4 nonzero displacement levels × 2 geometries), **486 (82.1%) showed resolved, substantive attenuation; 0 showed enrichment; 0 showed practical equivalence; and 106 (17.9%) were numerically non-convergent** (Table 2). Under the registered classifier this is unambiguously an **attenuation-only** outcome ($f_{att} = 0.821$, $f_{enr} = 0$). The 106 unresolved conditions are, without exception, non-convergent policy solves rather than converged-but-ambiguous effects: **among the 486 conditions whose numerics resolved, attenuation is unanimous (486/486).** No resolved condition showed the value of displaced feedback holding steady or increasing.

Displaced feedback that stays informative — provably non-reducible to a change in channel reliability (§3.4) — nonetheless loses substantial decision value when it evaluates the wrong target. This is a distinct effect from Gate 1's reliability axis, not a re-derivation of it.

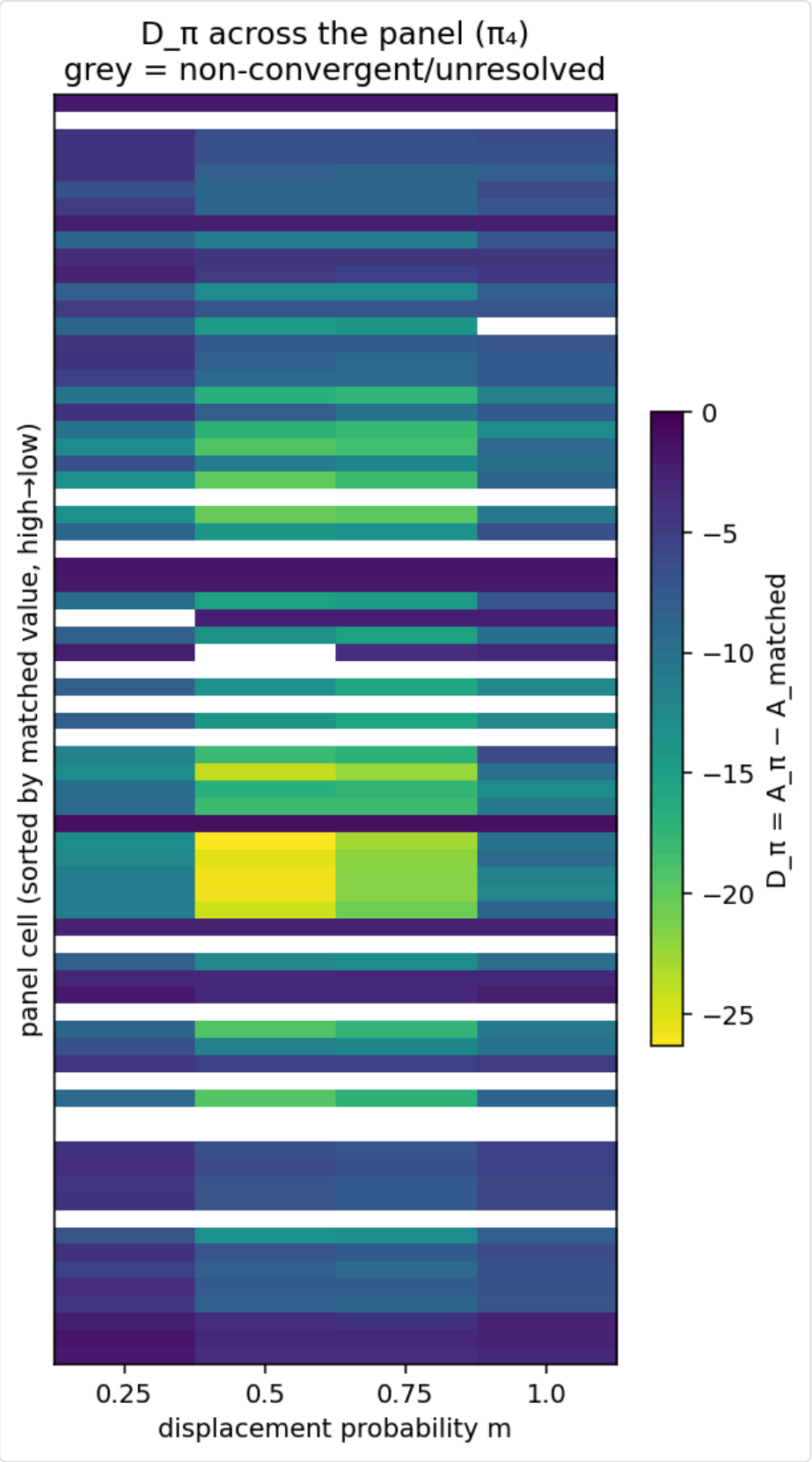
D_π across the panel (π₄)
grey = non-convergent/unresolved
panel cell (sorted by matched value, high→low)
0.25
0.5
0.75
1.0
displacement probability m
0
−5
−10
−15
−20
−25
D_π = A_π − A_matched

Figure 2. Signed advantage change D_π across the panel (π_4), cells ordered by matched value (high to low), columns the four displacement levels; grey marks non-convergent or unresolved conditions. Attenuation (D_π < 0) deepens with displacement across essentially the whole panel, with no resolved cell showing D_π ≥ 0.

Table 2. Outcome classification over the eligible set (E = 592 = 74 cells × 4 displacement levels × 2 geometries).

Class	Count	Fraction of E
Attenuation-substantive (D_π + ε_D ≤ −δ_min)	486	82.1%
Enrichment-substantive (D_π − ε_D ≥ +δ_min)	0	0%
Practically-equivalent (|D_π| + ε_D < δ_min)	0	0%
Effect-unresolved (converged, interval straddles δ_min)	0	0%
Numerically-unresolved-nonconvergent	106	17.9%
Among resolved conditions	**486 / 486 attenuation**	**100%**

δ_min = 0.05 · median(A over panel) = 0.648. Global outcome: **attenuation-only** (f_att = 0.821 ≥ 0.1, f_enr = 0).

The 106 non-convergent conditions are distributed evenly across displacement levels (26/26/24/30 at m = 0.25/0.5/0.75/1.0) and balanced across geometries (π_4 51, π_{2+2} 55), affecting 19 of the 74 cells. A plurality — 50 of 106 (47%) — fall in the two lowest cue-reliability rows (r = 0.25, 0.57): the rows where the *matched* Gate-1 baseline was itself least resolved. The displaced non-convergence thus concentrates where the underlying problem was already numerically hardest, not where displacement is largest — the by-m distribution (26/26/24/30 across m = 0.25/0.5/0.75/1.0) is nearly flat and the by-geometry split (π_4 51, π_{2+2} 55) is balanced. This is the signature of inherited difficulty rather than a displacement-specific artifact: the displaced solves fail to converge in the same region the matched solves strained, and evenly across the very parameter whose effect is under study. Per the frozen no-escalation discipline these conditions are censored, not pursued to finer grids, and they support no inference either way.

4.3 The value profile is non-monotone in displacement

The signed advantage change $D_{\pi}(m)$ is not monotone in m. Across the panel it descends to a minimum near m = 0.5 — the maximally ambiguous mixture, where the comparator is equally likely to evaluate a or π(a) — and partially recovers toward m = 1, where the channel becomes a clean reliability-q signal about the single displaced class π(a) and thus regains informativeness about a fixed (if wrong) target. For example, at cell (1,6,6,3) the 4-cycle gives D_{π} = −11.5, −24.4, −20.7, −8.6 at m = 0.25, 0.5, 0.75, 1.0 (Figure, right panel). Monotone decay was neither assumed nor observed; the recovery toward full displacement is a direct consequence of the mechanism's structure and was registered as a possible shape in advance.

4.4 The geometry of displacement matters, with a sign reversal

Geometry dependence is **established** by the registered criterion: the contrast $G(m) = D_{\pi_4} - D_{\pi_{2+2}}$ clears δ_{min} on five adjacent cell pairs, including three at m = 0.5 — $(1,5,6,3) \leftrightarrow (1,6,6,3)$, $(1,6,6,3) \leftrightarrow (2,6,6,3)$, and $(1,6,4,0) \leftrightarrow (1,6,4,1)$ — with the pattern replicating at m = 0.25 and m = 0.75. The effect is spatially coherent, not a scatter of isolated conditions.

Its structure is the paper's central finding. The sign of the geometry contrast **reverses with displacement probability** (Figure, left panel): among resolved conditions it is unanimously positive at low displacement (3 of 3 at m = 0.25; 6 of 6 at m = 0.5 — the 4-cycle preserving more value than the double-transposition), then unanimously negative at m = 0.75 (0 of 2 positive), and split at m = 1.0. The mechanism is visible in the paired $D_{\pi}(m)$ profiles: both geometries dip to a trough at m = 0.5 but the double-transposition dips deeper (π_4 −24.4 vs π_{2+2} −28.4 at cell (1,6,6,3)), and as displacement resolves toward a clean single-target channel the two profiles cross. Two displacements of *equal probability but different geometry* thus impose different — and, across the displacement range, oppositely ordered — costs on the value of feedback.

The value of an informative signal depends on the geometry of the latent distinction it reports, not only on how reliably or how often it reports. A signal can carry undiminished information and remain correctly modelled by an optimal controller, yet lose its worth because it partitions the hidden state relative to the wrong intervention target — and the shape of that loss depends on *which* wrong target, with a crossover as the displacement deepens.

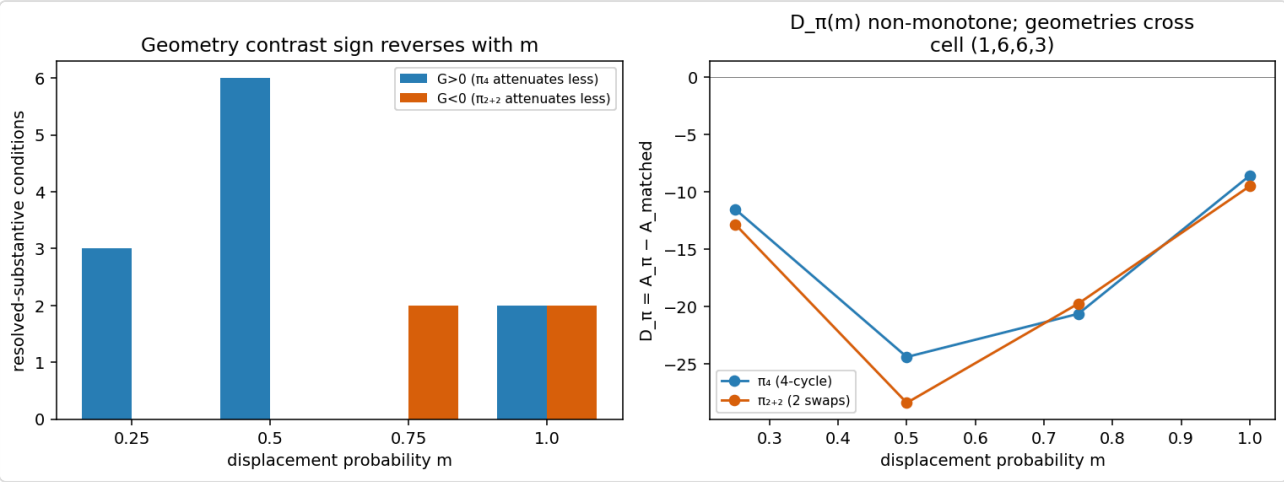

Figure 3. *(Left)* Sign of the geometry contrast G(m) = D_π_4 − D_π_{2+2} by displacement probability, counted over resolved-substantive conditions: unanimously positive at m = 0.25 and 0.5 (the four-cycle preserving more value), reversing to negative at m = 0.75, split at m = 1.0. *(Right)* Representative signed advantage-change profiles D_π(m) for cell (1,6,6,3), both geometries, showing the non-monotone trough at m = 0.5 and the crossover as displacement resolves toward a single-target channel. Data frozen in `es22_gate2_adjudicated.json`.

4.5 Enrichment was possible but did not occur

The design admitted enrichment — the Blackwell-incomparability of the displaced and matched channels means displaced feedback could in principle carry more decision-relevant information than matched feedback about some secondary latent, yielding A_π > A_matched. No resolved condition on this panel showed it. This is an honest negative on a registered possibility, not an outcome the design precluded: the incomparability that makes enrichment *possible* does not make it *occur* in this control problem under these losses.

5. Discussion

5.1 What the study establishes

Within one exactly-solved control problem, the decision value of a scalar feedback signal falls substantially when the signal evaluates a systematically displaced target — resolved, substantive attenuation in 82.1% of tested conditions, and unanimous attenuation among every condition whose numerics resolved. The effect is distinct from the reliability axis the model already contains: the displaced channel is provably not representable by any scalar reliability q (it induces a three-level likelihood structure, or at maximal ambiguity a class partition, that no two-level q-channel can produce), so this is not the requisite-variety story retold. And the effect has geometric structure: two displacements of equal probability but different permutation geometry impose different value costs, with the ordering reversing across the displacement range.

Taken together these support the paper's thesis — that requisite variety is necessary but not sufficient, and that a second constraint, requisite alignment, binds independently and has geometry. But the strength of that support is uneven across the three findings, and honesty requires separating them.

5.2 The marginal attenuation is partly confounded with information loss; the geometry effect is not

The clean version of the requisite-alignment claim would be: value falls under displacement *at constant information content*, so the loss cannot be an information-quantity effect in disguise. This holds for the geometry finding but not, without qualification, for the raw m-attenuation.

The displaced channel's marginal mutual information about the true class is not invariant in m. It equals the matched channel's at m = 0 and again at m = 1 (where the signal is a perfect-reliability report about the single class $\pi(a)$), but dips at interior m — to roughly 34–39% of the matched MI at m = 0.5, across the reliability range. So part of the attenuation as m rises from 0 toward 0.5 *coincides* with a genuine fall in how much the channel tells the controller about the state. A critic could reasonably say that this component of the effect is consistent with a variety/information account and does not, on its own, demonstrate a distinct alignment constraint.

Two things bear on this objection, and only one of them is settled.

First, the attenuation is disproportionate to and differently-shaped from the MI loss: MI is symmetric about m = 0.5 (it falls then recovers symmetrically), whereas the value profile, though also non-monotone, is not the image of the MI curve — value depends on how the surviving information maps onto the *action-relevant* distinction, not only on its quantity.

Second, the geometry contrast holds information content fixed in a specific and limited sense. The two registered geometries displace every action to exactly one other class, so their likelihood vectors are relabelings of a common multiset and their single-observation mutual information under a *uniform* prior is identical at every m (§3.4). It does not follow that they are informationally equivalent along a trajectory. At non-uniform beliefs the geometries assign the same likelihood levels to different latent classes and their per-observation MI diverges — by up to 0.22 bits at q = 0.85, m = 0.75, which exceeds the channel's entire uniform-prior information content in parts of the parameter range. The controller's belief is uniform only at t = 0. The geometry contrast is therefore conducted almost entirely at beliefs where the §3.4 equality does not hold.

An earlier version of this section drew the stronger inference — that the value difference between the geometries could not be an information-quantity effect, since the quantity was equal by construction. That inference is withdrawn. It relies on a single-observation, uniform-prior equality to license a claim about a sequential problem, and the equality does not survive the transition.

What the geometry contrast establishes without qualification is that the difference is not a *variety* effect. Output cardinality, reliability q, displacement probability m, and the likelihood multiset are identical across π_4 and π_{2+2} by construction, so no account in terms of how many distinctions the channel can make can separate them. This is the claim the requisite-alignment refinement of the Goodhart–Ashby synthesis requires — that requisite variety is necessary but not sufficient — and it is untouched by the correction. Note also that the paper does not need the stronger claim to be novel against value-of-information theory (§2.4): that information quantity underdetermines decision value is a settled starting point in that literature, and the contribution here was never the general claim but the solved, geometrically structured case.

Whether the difference is additionally free of any information-quantity explanation is now an empirical question rather than a construction, and it is registered as one.

> ***RP-XXVII-1 (registered; posterior to the value run, prior to this test).*** *Along the realized belief trajectories of run 16bc675b, accumulate per-step belief-conditioned mutual information separately for π_4 and π_{2+2} over each condition's horizon. Prediction: the accumulated-information gap does not reproduce the value gap — in particular it does not exhibit a sign reversal near m = 0.75. Falsification: if accumulated information differs between the geometries with the same sign as the value difference and reverses at the same displacement level, the geometry effect is not separable from an information-quantity account within this model, and the orthogonality claim is withdrawn rather than narrowed.*

The asymmetry of the test is worth stating. Marginal MI is symmetric about m = 0.5 while the value trough falls near m = 0.75; an information-quantity account of the geometry effect would have to explain a reversal at a displacement level where no information measure in this model reverses. A confirmed RP-XXVII-1 would therefore establish the dissociation by demonstration rather than by construction, which is a stronger result than the one being withdrawn.

The framing (§1–2) should therefore lead with the geometry result as the load-bearing evidence and present the m-attenuation as supporting but partially confounded, rather than treating all 82% as equally clean demonstration of an alignment-not-variety effect.

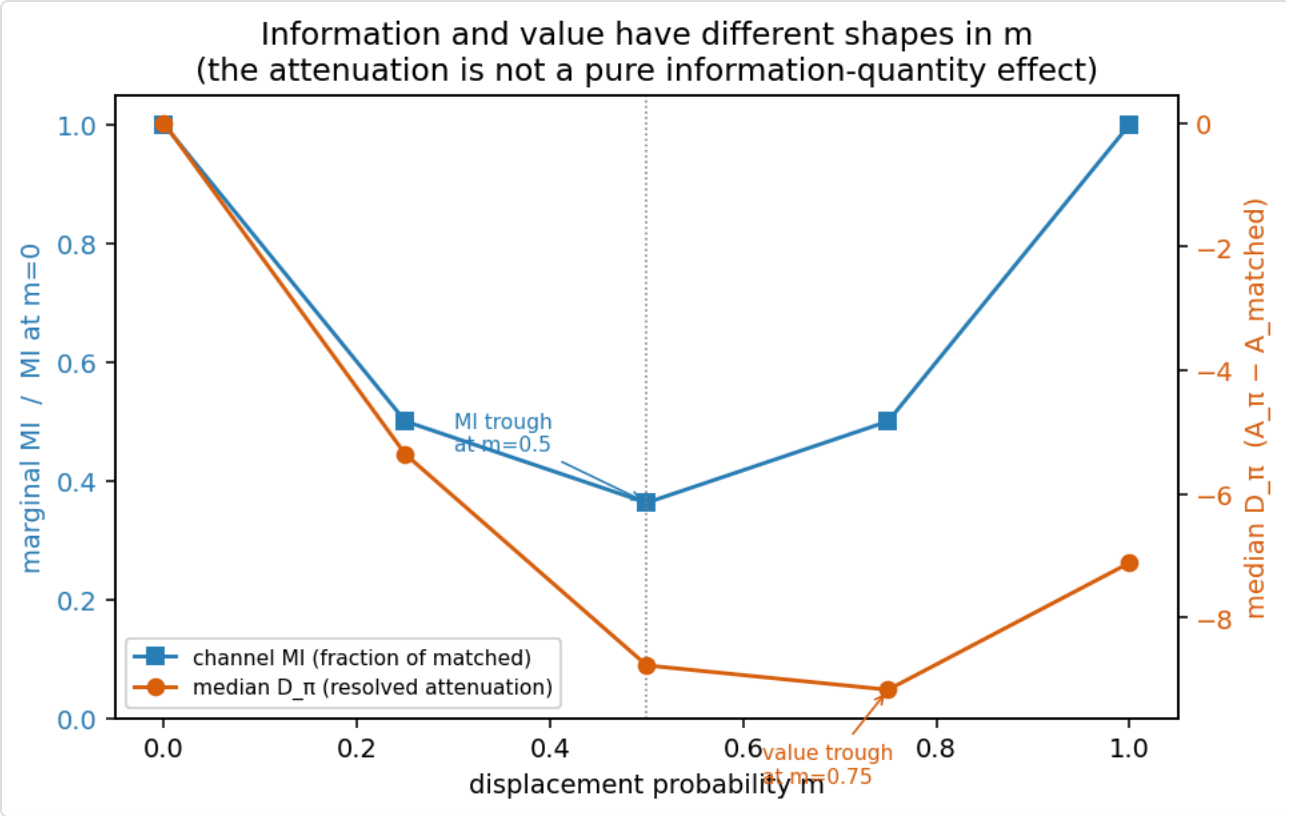

Figure 4. The channel's marginal mutual information (left axis, blue) falls symmetrically to a trough at m = 0.5 and recovers by m = 1, whereas the median resolved value cost D_π (right axis, orange) troughs later, near m = 0.75. The two curves have different shapes in m, so the value attenuation is not a simple image of the information loss — and the geometry contrast (Figure 3), which holds variety and uniform-prior MI exactly equal, isolates the component of the effect that no variety measure can explain. Whether any information measure can explain it is registered as an open test (RP-XXVII-1, §5.2).

5.3 What the study does not establish

Prevalence. The 82.1% is a fraction of tested conditions on a purposive 74-cell panel selected for existence and heterogeneity, stratified across the value magnitude but not sampled uniformly from the 1,121-cell active region. It is not an estimate of the proportion of that region that attenuates, and should never be reported as one. Whether attenuation is this pervasive across the full active region is an open question requiring the registered uniform-sample follow-up. The finding is existence, robustness, and geometric structure — not prevalence.

The censored conditions. Of the 592 conditions, 106 (17.9%) did not converge at the registered grid levels and are censored. They are not evidence of anything — not of attenuation, not of its absence. They cluster partly (47%) in the two lowest cue-reliability rows where the matched Gate-1

baseline itself sat near its resolution edge, so the non-convergence plausibly reflects inherited numerical difficulty rather than a feature of displacement; but "plausibly" is the operative word, and per the frozen no-escalation discipline we did not pursue them to finer grids to find out. A reader should treat the resolved-unanimous attenuation as strong within its 82% and make no inference about the censored 18%.

Enrichment. The design permitted the displaced signal to be worth *more* than matched feedback: the two channels are Blackwell-incomparable, so displaced feedback could in principle carry more decision-relevant information about some secondary latent, yielding $A_\pi > A_{matched}$. No resolved condition on this panel showed it. This is a genuine negative on a registered possibility, and it should be reported as such — the incomparability that makes enrichment *possible* does not make it *occur* under these losses and this parameterisation. It does not follow that enrichment cannot occur under other losses, other K, or other latent structure; only that this model, asked honestly, did not produce it.

Generality beyond the model. The result is a theorem-and-computation about one finite POMDP with $K = 4$, a specific loss structure, a binary feedback channel, and a known displacement. The requisite-alignment *concept* is proposed as general — it is a claim about the relationship between what an observable resolves and what an action depends on, which is not specific to this model — but the *evidence* is one solved instance. The paper should be read as establishing that the constraint is real and structured in at least one exactly-analysable case, not as measuring its magnitude in any governance system.

5.4 The non-monotonicity and the meaning of "maximal" misalignment

A feature worth dwelling on, because it is counterintuitive and because it disciplines the concept: the worst case is not maximal displacement. Value is lowest near $m = 0.5$ — where the signal is maximally *ambiguous* between the right target and a wrong one — and partially recovers at $m = 1$, where the signal becomes a perfectly reliable report about a *consistently* wrong target. A consistently wrong signal is more useful than an ambiguously-sometimes-right one, because a controller that knows the displacement can partially invert a consistent mapping but cannot disambiguate a mixture. This means "misalignment" is not a monotone scalar with a worst point at the extreme; its cost peaks at intermediate ambiguity. For the governance reading, the caution is against equating "maximally misaligned metric" with "maximally harmful metric" — a dashboard that is reliably measuring the wrong thing may be less damaging than one that is unpredictably measuring sometimes the right and sometimes the wrong thing, because the former can be corrected for and the latter cannot.

5.5 Relation to the deferred adaptive gate

Throughout, the controller *knows* the displacement — π and m are given, and it updates optimally under known ambiguity. This is deliberate: it isolates the value cost of misalignment-as-given from the separate problem of *inferring* the misalignment. The obvious next question — can a controller that does *not* know the displacement learn it, and at what cost — is a distinct study, and one this design specifically set aside. It is not a hidden limitation but a registered scope boundary: the adaptive version requires a controller maintaining and updating a belief over the displacement itself, which changes both the mechanism and the estimand (from value-of-misaligned-feedback to value-of-learning-the-misalignment). The present result is the necessary baseline for that study — one must know what misalignment costs when known before asking what it costs when it must be discovered.

5.6 Implication for the governance program

If requisite alignment is a real constraint independent of requisite variety, the series' prescription — raise the effective dimensionality of the observation channel — is necessary but incomplete. An institution can expand its dashboards along more independent dimensions, transmit those signals without compression, and still govern badly if the distinctions its metrics resolve are misaligned with the distinctions its interventions turn on. The dimensional prescription answers *how much* to observe; the alignment constraint asks *what* to observe — which distinction, matched to which lever. And because alignment has geometry, the corrective is not a scalar "measure the right thing" but a structural question about how the observed distinctions are arranged relative to the space of available interventions. This is, for the series, a second axis of architectural failure alongside dimensional insufficiency — one that a system can fall into precisely by succeeding at the first, adding rich, faithfully-transmitted, high-dimensional observation of the wrong distinctions. The solved model does not show that real governance systems fail this way; it shows that the failure mode is structurally real, geometrically structured, and invisible to a purely dimensional account.

References

Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall, London. [Law of Requisite Variety.]

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ.

Blackwell, D. (1951). Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 93–102. University of California Press, Berkeley.

Blackwell, D. (1953). Equivalent comparisons of experiments. *Annals of Mathematical Statistics*, 24(2), 265–272. doi:10.1214/aoms/1177729032.

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2), 89–97. doi:10.1080/00207727008920220.

Goodhart, C. A. E. (1975). Problems of monetary management: the U.K. experience. In *Papers in Monetary Economics*, Vol. I. Reserve Bank of Australia. [Goodhart's Law.]

Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1), 22–26. doi:10.1109/TSSC.1966.300074.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2), 99–134. doi:10.1016/S0004-3702(98)00023-X.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423 and 27(4), 623–656. doi:10.1002/j.1538-7305.1948.tb01338.x.

Series-internal references (Governance as Engineering): Paper I (the averaging problem), Paper II (the fractality requirement), Paper III (constitutional unobservability), Paper IV (commons collapse under low-dimensional channels), Paper VI (the Goodhart–Ashby synthesis), Paper VII (integration without compression). Full citations in the series bibliography at bjornkennethholmstrom.org.

Note on numerical methods. The simplex interpolation used by the solver is the CDF-Kuhn (Freudenthal) triangulation of the regular grid; the Gate-1 lower-bound certificate uses an empirical-Bernstein concentration bound. Both are standard and are described with implementation detail in the reproducibility artifacts (Appendix A) rather than claimed as contributions.

Appendix A. Reproducibility

All numerical results derive from a fixed chain of frozen artifacts. Each is content-addressed by the first 16 hex digits of its SHA-256; the run is deterministic (no RNG in the solver path; the one Monte-Carlo element, the Gate-1 lower-bound certificate, is a validation bound not used in any reported value). The Gate-2 production run carries run_id 16bc675b, bound to the panel, solver hash, grid levels (G = 40/54/64), horizon (H = 64), geometries, and displacement levels; a mismatch in any of these rejects the checkpoint.

Artifact	File	SHA-256 (16)
Model contract (ES-2.0 rev5)	`ES-2.0-model-contract.md`	2c71487c2ce18287
Gate-1 preregistration	`ES-2.1-gate1-prereg.md`	154d403eacde14ef
Gate-2 preregistration (rev1.1)	`ES-2.2-gate2-prereg.md`	ea7db6b5ff508a2e
Gate-1 solver (Kuhn)	`es21_gate1_solver_kuhn.py`	85560ac1e8c2d342
Gate-2 solver (Kuhn, displaced channel)	`es22_gate2_solver_kuhn.py`	53424370260e2656
Independent H = 2 validator	`es22_h2_exact.py`	1f42d5762eb1e89f
Gate-2 production driver	`es22_run_gate2.py`	f1585246503abaeb
Deterministic post-processing	`es22_postprocess_results.py`	5bdbe48637c4c600
Frozen 74-cell panel manifest	`es22_panel_manifest.json`	7a525ac3ca1cfe14
Gate-1 activation surface (amended)	`es21_step5_amended.json`	22d1fbf37ddb29d1
Gate-2 raw results (run 16bc675b)	`es22_gate2_results.json`	e98b42bd45ab685b
Gate-2 adjudicated re-analysis	`es22_gate2_adjudicated.json`	4211345f293f3f60

Solver. Belief-grid backward induction over the K = 4 simplex, CDF-Kuhn (Freudenthal) simplex interpolation at grid levels G = 40 → 54 → 64, horizon H = 64. The Gate-2 solver is a minimal diff from the Gate-1 solver, changing only the feedback observation likelihood to the three-level displaced channel (§3.4); all other machinery is identical, and the m = 0 identity (V1) confirms the reduction is exact.

Panel and classification. The 74-cell panel is selected deterministically from the frozen Gate-1 surface by the rule in §3.5 (qualified interior $A > 11 \cdot \varepsilon_A$, tertile stratification with SHA-256 hash ordering, registered anchors and slices where qualified, one adjacency pair per tertile) — no random sampling. Classification is the interval rule of §3.5 with $\delta_{min} = 0.05 \cdot \text{median}(A \text{ over panel}) = 0.648$, gated on independent convergence of both policy values (§3.6). The post-processing script regenerates every reported count, fraction, and the geometry adjudication from the raw results JSON without re-solving, and is the authority for the numbers in §4; the two driver-side reporting issues it corrects (pooling of the unresolved count; the permissive geometry flag) are documented in its header.

Validation matrices. The V1–V8 obligations (Table 1) were evaluated before any scientific condition was interpreted; the exact $H = 2$ enumeration (V7) shares no interpolation core with the production solver and independently confirms exact model conjugacy symmetry (spread 4×10^{-16}), establishing that the production solver's small conjugacy spread is interpolation bias within the registered envelope, not a model asymmetry.

Hardware. Ryzen 7 3700X, 16 threads; the production run completed in roughly 2.5 hours at 12 worker processes. numpy/scipy only; no GPU, no network in the solve path.
